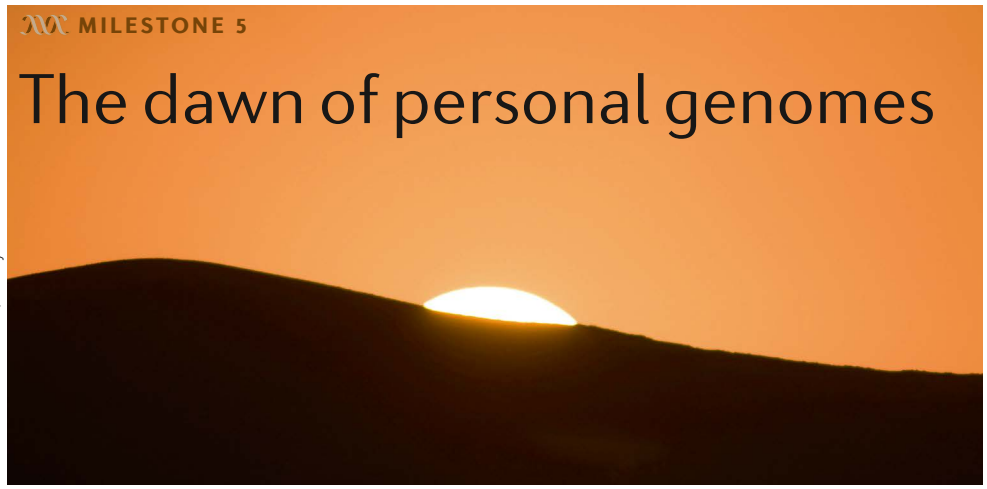
MILESTONE 5

The dawn of personal genomes



Credit: Redmond Durrell / Alamy Stock Photo

The Human Genome Project showed how large-scale international collaboration could generate a resource of enduring scientific value: a human genome reference sequence (MILESTONE 1). But with a cost of approximately US$300 million and requiring several years, technological advances would be required for whole-genome sequencing to become more widely achievable. The year 2008 saw the publication of pioneering applications of next-generation sequencing technologies to generate genomes of human individuals at a fraction of the cost and time of Sanger sequencing.

In *Nature*, Bentley et al. and Wang et al. reported the genomes of an African individual and an Asian individual, respectively. Bentley et al. introduced a novel massively parallel reversible terminator approach, which was adopted by Wang et al. as well as in a separate study by Ley and colleagues (MILESTONE 6). Originally known as Solexa sequencing, this technique remains the mainstay of short-read Illumina sequencing approaches to the present day.

Whereas Sanger sequencing determines the sequence of a single DNA clone per capillary channel, the major advancement of next-generation approaches was to enable the sequencing of millions of DNA clones simultaneously. To achieve this parallelization, single DNA molecules were immobilized on a flow-cell surface, followed by localized amplification to generate a focal cluster of identical DNA molecules from each starting molecule. A camera then images the flow cell as the DNA clones undergo sequential rounds of controlled, stepwise single-nucleotide addition, with each of the four possible added nucleotides labelled with a different coloured fluorophore that is removed after each cycle. For each clonal spot on the flow cell, the sequential colour changes reveal the DNA sequence.

Although the read lengths of ~35 bases achieved at the time were short relative to Sanger sequencing (and indeed relative to current-day improved short-read methods and long-read methods), the high depth of coverage (>30×) allowed reads to be overlapped and mapped to the existing reference genome, thus generating genome sequences that were near complete except for challenging repetitive regions. Each genome sequence was produced for less than US$500,000 in a few weeks. This progress represented a major step change in affordability and set the stage for time and cost reductions that continued until recently.

New human genomes enable analyses of genetic variation between individuals, and sequencing is particularly powerful for identifying novel variation relative to genotyping microarrays, which require the variants to be already known and predesigned onto the genotyping chip. Bentley et al. and Wang et al. identified several million single-nucleotide variants (SNVs) relative to the human reference genome, many of which were novel.

Despite the short reads, both studies identified thousands of larger structural variants. This was possible owing to the paired-end nature of the sequencing whereby the ends of DNA molecules are sequenced but an intervening genomic region of known approximate length remains unsequenced. Structural variation can be identified when the two sequenced ends map to the reference genome at unexpected distances or orientations to each other. Both studies highlighted the prevalence of polymorphisms in transposable element insertion sites as a major source of human genetic variation.

> “major groundwork for future larger-scale human sequencing projects”

These studies laid major groundwork for future larger-scale human sequencing projects (MILESTONE 13), not just by demonstrating the feasibility and resource value of human genome sequences but also in testing the suitability of associated bioinformatic tools and design considerations such as necessary sequencing depths per individual.

Such studies emerged at a time of great interest in possibilities for personal genomics. As the human reference genome is a composite of genomes from several anonymous donors, it does not represent any single individual. Opening up genome sequencing to individuals allows participants to be informed about not just their own ancestral make-up but also any potential disease-associated genetic variants that may inform risk of future diseases for themselves or their family members.

Lower-depth (~7×) personal genomes of two notable named scientists were revealed at a similar time by Levy et al. using Sanger sequencing and Wheeler et al. using the (now-discontinued) 454 pyrosequencing system. Collectively, these four personal genomes identified several genetic variants of potential medical relevance for the individuals.

Overall, these studies paved the way for the widespread sequencing that we see today in population genomics projects and in clinical applications.

Darren J. Burgess,
Nature Reviews Genetics

ORIGINAL ARTICLES Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008) | Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008)
FURTHER READING Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007) | Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008) | Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016)